The Sustainable AI Paradox

What You Need to Know About AI's Environmental Footprint

Adam Wynne | adam@wynnetech.ai
March 23, 2026

Executive Summary

Sustainability leaders are being asked to adopt AI: to automate ESG reporting, track emissions, monitor suppliers, and accelerate compliance. The tension is obvious. Using AI to reduce environmental impact when AI itself has an environmental footprint.

I cover four things in this paper: AI's actual energy, water, and carbon impact; what the major vendors actually disclose and what that means for your reporting; five practical steps to manage AI's footprint responsibly; and what the best available evidence says about whether AI's carbon cost is worth it. A technical appendix covers the full tooling landscape for IT teams and cloud architects.

The core finding, drawn from IEA's independent analysis, is that AI applications have the potential to reduce more carbon than they create. AI's emissions reduction potential in end-use sectors reaches 1,400 Mt CO2 by 2035. AI's own data center emissions over the same period: 300-500 Mt. The reductions are 3-4x larger than AI's own footprint, but only with deliberate adoption.

---

1. The Reporting Gap

When you deploy AI to automate Scope 3 data collection, a reasonable question follows: what is the Scope 2 emissions cost of running that system?

Most AI vendors don't make this easy to answer. Most sustainability frameworks don't yet require it. Most organizations deploying AI for ESG purposes haven't asked the question formally.

The sustainability community needs to push AI vendors for standardized disclosure and build the reporting infrastructure to analyze it. We wouldn't accept this opacity from an energy supplier. We shouldn't accept it from an AI provider either.

---

2. The Actual Impact: Energy, Water, and Carbon

2.1 Energy

A single ChatGPT query consumes roughly 0.34 Wh, about what a high-efficiency lightbulb uses in a couple of minutes.[1] Complex reasoning queries, the kind increasingly used for document analysis and report drafting, can use up to 50x more.

At the infrastructure level, the numbers are considerably larger. Global data centers consumed 415 TWh in 2024, roughly 1.5% of global electricity. The IEA projects this to double by 2030, with US data centers already exceeding 4% of the national energy mix.[2]

There are two phases to AI energy consumption that sustainability leaders need to understand. **Training** is the one-time process of building a model, computationally intensive but it happens once. **Inference** is what happens every time someone uses it: a query goes in, a response comes out, compute runs. It scales directly with adoption. GPT-4's training required an estimated 50 GWh,[3] a large number, but a fixed one. Inference has no ceiling.

That distinction matters because Deloitte projects inference will account for roughly 66% of AI energy consumption by 2026, potentially reaching 80% long-term.[4] Every new user, every new query, adds to the load. Usage-based tracking is more relevant than model-level energy estimates.

2.2 Water

AI workloads require significant cooling, and cooling requires water. UC Riverside researchers found that GPT-3 consumed roughly 500ml of water per 10-50 medium-length responses, depending on where it was deployed.[5] Newer models are considerably more efficient: Google disclosed 0.26ml per Gemini query in 2025. But aggregate consumption at scale remains significant. Training GPT-3 required approximately 700,000 liters. Looking ahead, US data center cooling is projected to need between 697 million and 1.45 billion additional gallons per day of peak water capacity.[6]

Water stress is increasingly a material ESG risk for data center operators. For organizations in water-stressed regions, or with water reduction commitments, this warrants explicit consideration in AI procurement decisions. It receives far less attention than energy and carbon.

2.3 Carbon

The carbon impact of AI depends enormously on where workloads run. BLOOM, trained on France's nuclear-heavy grid, produced 25 tCO2e. GPT-3, trained on the US average grid, produced 502 tCO2e.[7] That is a 20x difference for the same type of activity, driven entirely by grid carbon intensity.

For individual users, the numbers are more modest. Ten queries per day for a year produces roughly 11 kg CO2. One transatlantic flight produces 1,000-2,000 kg CO2. The individual footprint is small. The aggregate footprint of enterprise-scale deployment is not.

Where your AI workloads run matters as much as how much you run. Grid carbon intensity is a controllable variable, and one that most organizations currently ignore.

---

3. What the Vendors Actually Tell You

To understand what vendors actually disclose, you need to look at two distinct layers: the AI platforms you interact with directly, and the cloud infrastructure those platforms run on. Most organizations use both. The reporting picture differs significantly between them, and the distinction determines what you can actually include in your ESG disclosures today.

3.1 AI Platform Layer

This is the layer most sustainability leaders interact with daily: Gemini, ChatGPT, Claude. It is also where disclosure is weakest. One provider leads clearly. The other two have significant ground to cover.

Google / Gemini — leading the field. In August 2025, Google published the most comprehensive first-party AI environmental disclosure in the industry to date. The median Gemini Apps text prompt uses 0.24 Wh of energy, emits 0.03 gCO2e, and consumes 0.26 milliliters of water.[8] Over a 12-month period, energy per prompt dropped 33x and carbon per prompt dropped 44x.[9] The methodology covers the full serving stack: active accelerator power, host CPU and memory, idle capacity, data center overhead, and embodied hardware. Independent analysts have noted legitimate limitations: Google uses market-based rather than location-based carbon accounting, and model training costs are excluded.[10] Both are worth noting. Neither undermines the value of having a formal, published methodology. No other platform comes close.

OpenAI / ChatGPT — no formal reporting. As of March 2026, OpenAI has published no standalone sustainability report and no verified Scope 1, 2, or 3 emissions figures.[11] The

primary disclosure is a CEO blog post stating the average ChatGPT query uses approximately 0.34 Wh and roughly one-fifteenth of a teaspoon of water.[12] That is a single executive statement, not an auditable reporting framework. No per-query carbon dashboard exists for ChatGPT or API users.

Anthropic / Claude — offsets but no disclosure. Anthropic states in the Claude 3 model card that it fully offsets its operational carbon emissions annually through verified carbon credits, working with external experts to measure its company-wide footprint.[13] That process requires internal carbon accounting — the data exists. What doesn't exist is any public disclosure of it. No Scope 1, 2, or 3 figures have been filed as of March 2026, and no per-query energy or carbon data has been shared with users.[14] One independent eco-efficiency analysis ranked Claude 3.7 Sonnet highest among major models for balancing reasoning performance with infrastructure efficiency, scoring 0.886 out of 1.0.[15] A useful signal on relative efficiency, but it does not substitute for formal emissions disclosure.

3.2 Cloud Infrastructure Layer

The three major cloud providers — Google Cloud, Azure, and AWS — all offer carbon reporting at the infrastructure level, but with meaningful differences in quality and accessibility.

Google Cloud. The most detailed of the three. BigQuery integration gives organizations granular exports for custom analysis, and Google's broader environmental reporting covers Scopes 1, 2, and 3 with market-based and location-based emissions data. For sustainability teams that need depth, it is the strongest option.

Azure Carbon Optimization. Azure's current carbon reporting tool (Azure Emissions Insights was deprecated in August 2025 and replaced by Azure Carbon Optimization) covers Scopes 1, 2, and 3 and breaks emissions down by service category (Compute, Network, and Storage) and by region. Accessible via the Azure portal and exportable via REST API or CSV. The important limitation: service categories do not distinguish workload purpose. AI inference, a database, and a web server all appear as "Compute." For organizations using OpenAI via their own Azure API accounts, this is the closest available proxy for AI compute emissions, but correlation with known AI adoption timelines and FinOps spend data is required to make the estimate meaningful.

AWS. The weakest of the three on reporting. Quarterly delays mean the data is rarely current enough for live ESG workflows, and CSV-only exports require manual processing. AWS has signaled improvement is coming, but as of March 2026 the tooling lags meaningfully behind its peers.

A critical caveat applies to all three: none of this reporting is automatic or universally accessible. Organizations using ChatGPT.com or Claude.ai directly have no visibility into the underlying cloud infrastructure at all. That layer is simply invisible to platform users. For API customers

running workloads on their own cloud accounts, infrastructure reporting requires the right subscription tier, enabled reporting tools, and deliberate configuration. It is a proxy, not a solution.

3.3 What This Means for Your Reporting

The gap is not just a transparency issue. It is a practical reporting problem. As of March 2026, no AI platform provides the data needed to include AI-specific emissions in Scope 2 or Scope 3 disclosures with precision. Platform users are largely in the dark. Even API customers with access to cloud infrastructure dashboards are working with imprecise proxies.

That will begin to change. The EU AI Act's energy disclosure requirements, effective August 2026, will require general-purpose AI providers operating in European markets to document energy consumption.[16] The Green Software Foundation's SCI for AI standard, ratified in December 2025, provides the measurement framework providers should be aligning to now.[17] Both create real pressure for disclosure to improve over the next 12-18 months. In the meantime, demanding this data in vendor procurement conversations is not unreasonable.

	Per-Query Disclosure	**Scope Emissions Filed**	**Infrastructure Dashboard**
Gemini (Google)	Full methodology (Aug 2025)	Yes	Yes, BigQuery export
ChatGPT (OpenAI)	CEO statement only, no formal reporting	No	No, Azure not accessible to platform users
Claude (Anthropic)	No	No	No, AWS not accessible to platform users

---

4. Five Practical Steps

These are sequenced by maturity. Start at step one. For IT teams and cloud architects looking for specific tools, dashboards, and technical implementation guidance, see the Appendix.

Step 1: Build a baseline from what you already have. Cloud provider carbon dashboards (Azure Carbon Optimization, Google Cloud Carbon Footprint, and AWS Customer Carbon Footprint Tool) report total compute emissions broken down by service category, not by workload purpose. You cannot isolate AI inference from a database or a web server in these

dashboards. What you can do is correlate. If your organization launched a major AI initiative in Q3 and total compute emissions jumped 40% in Q3, that correlation is defensible even without per-workload tagging. Pair it with FinOps spend data tagged to AI projects and you have a two-source estimate: total compute trends from the cloud dashboard, AI's share of compute spend from FinOps. Imprecise, but a reasonable methodology that most ESG frameworks will accept while industry standards mature.

Step 2: Ask your IT team to right-size AI models. Not every task requires a frontier model. Working with your IT or data team to match model size to task complexity, using smaller models for routine work like document summarization, data extraction, and report drafting, can reduce AI energy consumption meaningfully. This is a technical decision, but the business direction comes from you: define what "good enough" looks like for each use case.

Step 3: Ask where your AI workloads run. Grid carbon intensity varies 10-20x across regions. Where your AI workloads run matters as much as how much you run. Ask your cloud or IT team whether your organization's AI compute is region-optimized for carbon intensity. Case studies report 5-20% emissions reductions from carbon-aware workload placement alone.[18] This is low-cost, high-leverage, and entirely within reach for most organizations.

Step 4: Treat inference optimization as a sustainability lever. Inference dominates total AI energy at scale. Efficient prompting, response caching, and using the smallest model that meets requirements are not just cost-saving measures; they are emissions reduction strategies. Brief your IT team on this framing. FinOps teams (cloud financial operations: the practice of managing and optimizing cloud infrastructure costs) are already managing inference costs. They may just not be connecting it to your sustainability reporting yet.

Step 5: Make FinOps your bridge to carbon reporting. If your organization has a FinOps practice (tracking cloud infrastructure spend by service, project, or resource tag) you can identify AI compute as a subset and use it as the foundation for AI carbon tracking today. AI spend is a direct proxy for AI compute, which is a direct proxy for AI energy. Map spend to energy using provider efficiency benchmarks, apply grid carbon intensity factors for the regions where workloads run, and you have a defensible AI carbon estimate sufficient for internal reporting and ESG disclosure drafting. Organizations without a FinOps practice should start one: the sustainability case is as strong as the cost case. Adopt the SCI for AI framework as your measurement standard as it matures,[17] and use the EU AI Act's August 2026 effective date as a near-term reporting milestone.[16]

---

5. Putting AI's Carbon Footprint in Perspective

The most authoritative independent analysis available, from the IEA, concludes that the emissions AI enables organizations to avoid will outweigh AI's own carbon footprint. AI's

emissions reduction potential in end-use sectors reaches 1,400 Mt CO2 by 2035. AI's own data center emissions over the same period: 300-500 Mt. The reductions are 3-4x larger than AI's own footprint.[19]

That is a projection. The production results are already coming in, and they point in the same direction. DeepMind's data center cooling optimization reduced cooling energy by 40%.[20] BrainBox AI, deployed across 600 Dollar Tree stores, saved 7.98 million kWh in a single year.[21] DeepMind's GNoME identified 2.2 million new crystal structures relevant to batteries, solar cells, and carbon capture.[22]

The question is not whether AI has an environmental cost. It does. The question is whether you deploy it deliberately, track its impact, and direct it toward outcomes that exceed that cost. The evidence says the potential is there. Whether you capture it depends on intention.

About the Author

Adam Wynne is the founder of Wynne Technologies, where he advises on company-wide AI technical strategy and helps product teams ship faster by combining product discipline with AI-accelerated development. After 20 years building SaaS products for organizations including Bosch, Armada Supply Chain Solutions, and the US Department of Energy's Pacific Northwest National Laboratory (PNNL), he now applies that expertise to climate challenges and sustainability technology.

Learn more at wynnetech.ai.

Footnotes

[1] Sam Altman, OpenAI blog post, 2025. Via MIT Technology Review, "The carbon footprint of AI," May 2025.

[2] IEA, "Electricity 2025: Analysis and forecast to 2027," January 2025. https://www.iea.org/reports/electricity-2025

[3] Epoch AI, "Training compute of frontier AI models," 2025.

[4] Deloitte, "Technology, Media & Telecommunications Predictions 2025."

[5] Li, P., Yang, J., Islam, M.A., Ren, S. "Making AI Less Thirsty: Uncovering and Addressing the Secret Water Footprint of AI Models," Communications of the ACM, 2023. GPT-3 consumes 500ml of water per 10-50 medium-length responses depending on deployment location. https://arxiv.org/abs/2304.03271

[6] UC Riverside & Caltech, March 2026.

[7] Luccioni, A.S., Viguier, S., Ligozat, A. "Estimating the Carbon Footprint of BLOOM," 2023. https://arxiv.org/abs/2211.02001. Patterson et al., "Carbon Emissions and Large Neural Network Training," 2021.

[8] Google Cloud Blog, "Measuring the environmental impact of AI inference," August 2025. https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference/

[9] Google technical paper, "Measuring the environmental impact of delivering AI at Google Scale," August 2025. https://services.google.com/fh/files/misc/measuring_the_environmental_impact_of_delivering_ai_at_google_scale.pdf

[10] Towards Data Science, "Is Google's Reveal of Gemini's Impact Progress or Greenwashing?", August 2025. https://towardsdatascience.com/is-googles-reveal-of-geminis-impact-progress-or-greenwashing/

[11] The Sustainable Innovation, "OpenAI Sustainability," March 2026. https://thesustainableinnovation.com/open-ai/

[12] Sam Altman, OpenAI blog post, 2025. Via Towards Data Science analysis, August 2025.

[13] Anthropic, Claude 3 Model Card, 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[14] Earth911, "Your AI Carbon Footprint: What Every Query Really Costs," March 2026. https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

[15] Earth911, "Your AI Carbon Footprint: What Every Query Really Costs," March 2026. Cites Jegham eco-efficiency analysis ranking Claude 3.7 Sonnet highest (score: 0.886) among major models. https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

[16] EU AI Act, Article 53(1)(e). Effective August 2, 2026.

[17] Green Software Foundation, "SCI for AI," December 2025. https://sci-for-ai.greensoftware.foundation/

[18] WattTime, Electricity Maps, and Green Software Foundation case studies.

[19] IEA, "The energy impact of AI," 2025. https://www.iea.org/topics/artificial-intelligence

[20] DeepMind blog, "AI for Data Centre Efficiency," 2016/2018. https://deepmind.google/discover/blog/

[21] BrainBox AI case study, 2025.

[22] Merchant, A. et al., "Scaling deep learning for materials discovery," Nature, 2023. https://www.nature.com/articles/s41586-023-06735-9

Appendix: Tools and Technologies for IT, Cloud Architecture, and Data Teams

This appendix is intended for the technical teams responsible for implementing AI carbon tracking, optimizing infrastructure, and building the reporting pipelines that sustainability leaders need. It covers four categories: emissions measurement tools, grid carbon intensity data sources, Kubernetes-native scheduling tools, and standards and frameworks.

A.1 Emissions Measurement Tools

CodeCarbon An open-source Python library that estimates carbon emissions from code running on a computer or server. It tracks CPU, GPU, and RAM energy consumption, applies the carbon intensity of the local grid, and produces a CO2e estimate per run. Most useful for organizations running AI workloads on their own infrastructure or cloud accounts. For third-party AI platforms, CodeCarbon is only available if the vendor has chosen to instrument it into their product — it cannot be deployed against external APIs. https://codecarbon.io/

Kepler (Kubernetes-based Efficient Power Level Exporter) A CNCF open-source project and Prometheus exporter that measures energy consumption at the container, pod, and node level in Kubernetes clusters. Provides granular real-time power metrics for organizations running AI workloads on Kubernetes, enabling per-workload carbon attribution. Actively developed under the CNCF Environmental Sustainability Technical Advisory Group (TAG ENV). https://github.com/sustainable-computing-io/kepler

Cloud Carbon Footprint An open-source tool that pulls usage data from AWS, Azure, and Google Cloud APIs and estimates carbon emissions per workload using published cloud provider efficiency data. More accessible than CodeCarbon for organizations already on public cloud, as it does not require code instrumentation. Outputs can feed into ESG dashboards. https://www.cloudcarbonfootprint.org/

ML CO2 Impact Calculator A web-based tool for estimating the carbon footprint of machine learning training runs. Useful for one-off estimates and benchmarking training jobs across hardware and regions. https://mlco2.github.io/impact/

TokenPowerBench The first open-source benchmark for LLM inference power consumption, released December 2025. Measures energy per token across model sizes. Key finding: energy per token scales roughly 7.3x from 1B to 70B parameter models (sub-linear), and Mixture-of-Experts (MoE) architectures use approximately one-third the energy of dense models

at similar quality levels. Useful for comparing model efficiency before procurement or deployment decisions. https://arxiv.org/abs/2512.03024

A.2 Grid Carbon Intensity Data Sources

These APIs provide real-time and forecast carbon intensity data for electricity grids by region. They are the data layer that powers most carbon-aware scheduling tools.

WattTime Provides real-time and forecast marginal emissions data for electricity grids across the US and internationally. Used by several cloud providers and scheduling tools as a primary data source. https://www.watttime.org/

Electricity Maps (formerly Tomorrow) Real-time and historical carbon intensity data for 160+ regions globally. Offers both a free tier and commercial API. More internationally comprehensive than WattTime for non-US deployments. https://www.electricitymaps.com/

Climatiq A carbon calculation API that covers emissions factors across cloud compute, energy, transport, and materials. Useful for organizations building custom carbon accounting pipelines that span AI and non-AI workloads. https://www.climatiq.io/

Green Software Foundation Carbon Aware SDK An open-source SDK that abstracts grid carbon intensity data from multiple providers (WattTime, Electricity Maps, and others) into a single interface. The foundation layer for many Kubernetes and cloud-native carbon-aware tools. https://github.com/Green-Software-Foundation/carbon-aware-sdk

---

A.3 Kubernetes-Native Carbon-Aware Scheduling

For organizations running AI workloads on Kubernetes, this tooling layer enables carbon-aware scheduling without modifying application code. Workloads are automatically shifted to lower-carbon time windows or regions based on real-time grid data.

Carbon Aware KEDA Operator (Microsoft/Azure) A Kubernetes operator that integrates with KEDA (Kubernetes Event-Driven Autoscaler) to scale workloads based on real-time carbon intensity. Sets a dynamic ceiling on maxReplicaCount during high-carbon-intensity periods, reducing scale-out when the grid is dirty and allowing full scaling when it is clean. Uses WattTime, Electricity Maps, or other providers as data sources. No application code changes required. Best suited for batch processing, ML training jobs, data analytics, and other time-flexible workloads. https://github.com/Azure/carbon-aware-keda-operator

Compute Gardener A Kubernetes scheduler that enables carbon-aware workload scheduling at the cluster level without any changes to application code or configuration. Workloads opt in via a simple scheduler annotation. Integrates with Electricity Maps and WattTime for real-time grid data. https://computegardener.io/

KubeGreen A Kubernetes operator focused on reducing energy consumption during off-hours by suspending idle workloads and namespaces on a defined schedule. Complementary to carbon-aware scheduling tools: handles baseline idle reduction while KEDA-based tools handle dynamic load shifting. https://kube-green.dev/

Kepler + Prometheus + Grafana The standard observability stack for Kubernetes energy monitoring. Kepler provides per-pod power metrics, Prometheus collects and stores them, and Grafana visualizes them. Enables teams to identify energy-intensive workloads, track efficiency improvements over time, and produce carbon attribution reports for sustainability teams. Setup guidance available via the CNCF TAG ENV working group.

A.4 Standards and Frameworks

SCI for AI — Green Software Foundation The first consensus-based standard for measuring AI carbon intensity, ratified December 2025. Extends ISO/IEC 21031:2024 and defines separate Provider scores (training and deployment) and Consumer scores (inference). The measurement framework AI vendors should be aligning to, and the standard sustainability teams should reference when requesting vendor disclosures. https://sci-for-ai.greensoftware.foundation/

FinOps for AI — FinOps Foundation AI compute cost management is the FinOps Foundation's current top priority. A certified FinOps for AI qualification launched March 2026. The connection to sustainability reporting is direct: compute cost is a proxy for compute energy, which is a proxy for compute carbon. Organizations with mature FinOps practices for AI are halfway to AI carbon tracking. The missing step is mapping spend to energy to carbon using provider efficiency benchmarks and grid intensity factors.
https://finops.org/wg/finops-for-ai-overview/

EU AI Act — Energy Disclosure Requirements Article 53(1)(e) of the EU AI Act, effective August 2, 2026, requires general-purpose AI providers operating in European markets to document energy consumption. This is the most significant near-term regulatory driver for vendor disclosure improvement. Organizations procuring AI for European operations should begin asking vendors for EU AI Act compliance timelines now.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689